Having briefly looked at all the data following observations were made regarding the tidiness & quality issues that need to be addressed in the data.

Tidiness:

1. 'rating_denominator','expanded_urls' does not provide any significance in 'twitter-archive-enhanced.csv'.
    a. Denominator ratings are all common for all tweets, hence no information can be retrieved.
    b. Expanded URLs provide no value as we already have all the data.
2. Columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_timestamp' all provide redundant information in 'twitter-archive-enhanced.csv'.
3. Removed duplicate jpeg url from 'image-predictions.tsv'.
4. Combining dog breed from image-predictions.tsv into master database.
    a. Dog breed predicted by neural net were imported into 'twitter-archive-enhanced.csv'.
5. Imported favourite and retweet count into master database from tweepy for easy visualization.
    a. Using 'Tweepy' API, favourite and retweet count were extracted from JSON for the 'tweet_id' in 'twitter-archive-enhanced.csv'and used
6. Stages of dog(doggo, floofer etc) mentioned in twitter_archive_enhanced.csv have been changed from four columns into one

Quality:

1. Sources of tweet can be modified into one word.
    a. Instead of entire text which is hard to interpret , one word such as iPhone or web can be written.
2. Removed rows where no dog predictions (p1_dog & p2_dog & p3_dog are all False) were there from 'image-predictions.tsv'.
    a. Since we are only concerned with dogs , I removed rows where the neural net predicted no dogs in all three cases.
    b. If at least one dog was predicted regardless of the confidence, that row was retained.
3. Selecting the predicted dog breed with strongest confidence from image-predictions.tsv and removed others from 'image-predictions.tsv'.

    a. Since we have multiple ‘dog_breed’ columns which are predicted , I selected the one which was correctly predicted as dog and had the highest confidence.
4. Removed retweeted tweets.
    a. Retweeted tweets provide no valuable information and hence removed.
5. Correcting numerator rating, i.e. removed ratings >14 & < 1.
    a. Numerators usually have a range from 1-14, high ratings are unusualy, It can be any other number mentioned in the tweet, hence these were removed.
6. Fixed the names of dog, removed names starting with lower case letter.
    a. Names column of dogs contained irregular data, names usually start with uppercase letter. Those names starting with lower case letter were removed.
7. Removed rows with denominator rating other than 10.
8. Removed Deleted Tweets.
    a. Deleted Tweets have Retweet_count and favorite_count as Zero